Team Name - Votrix

Team Members - Mohit Kumar Rathod (MT22041), Bhaswanth Gudimella (MT2025), Rohitkumar Tangudu (MT22060), Saketh Ragirolla (2021092)

Beijing Multi-Site Air-Quality Data - Dataset Description Report

Introduction

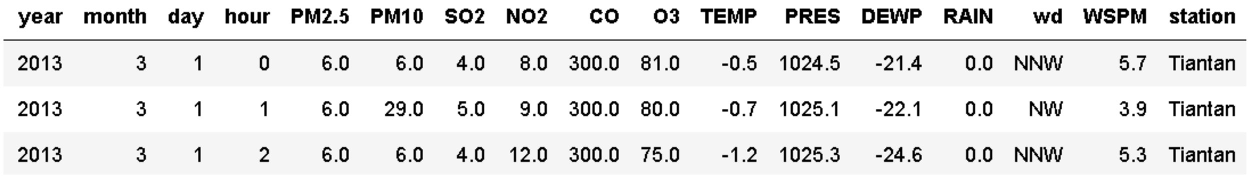
In this report, we will describe the "Beijing Multi-Site Air-Quality Data" dataset. This dataset contains valuable information about air quality in Beijing, China, collected from multiple monitoring sites over a period of several years. Understanding air quality is crucial for addressing environmental and public health concerns, making this dataset a valuable resource for researchers and policymakers.

Dataset Overview

Dataset Name: Beijing Multi-Site Air-Quality Data

Data Source: UCI Machine Learning Repository

- **Data URL**: https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data



year	month	day	hour	PM2.5	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	wd	WSPM	station
2013	3	1	0	6.0	6.0	4.0	8.0	300.0	81.0	-0.5	1024.5	-21.4	0.0	NNW	5.7	Tiantan
2013	3	1	1	6.0	29.0	5.0	9.0	300.0	80.0	-0.7	1025.1	-22.1	0.0	NW	3.9	Tiantan
2013	3	1	2	6.0	6.0	4.0	12.0	300.0	75.0	-1.2	1025.3	-24.6	0.0	NNW	5.3	Tiantan

figure-1: 3 sample points

Data Description

The Beijing Multi-Site Air-Quality dataset provides a comprehensive set of air quality measurements, meteorological data, and other relevant information collected from multiple monitoring sites across Beijing. This dataset aims to understand the air quality in different parts of the city and its correlation with various meteorological factors.

Key Features

1. **Site**: A categorical attribute representing the monitoring site within Beijing where the data was collected. There are several sites, each identified by a unique code.

2. **Date**: The date of data collection. It is presented in the format YYYY-MM-DD.

3. **Hour**: The hour of the day when the data was recorded. It is presented in a 24-hour format.

4. **PM2.5**: Concentration of particulate matter with a diameter of 2.5 micrometers or smaller (µg/m^3).

5. **PM10**: Concentration of particulate matter with a diameter of 10 micrometers or smaller (µg/m^3).

6. **NO2**: Concentration of nitrogen dioxide (µg/m^3).

7. **CO**: Concentration of carbon monoxide (mg/m^3).

8. **O3**: Concentration of ozone (µg/m^3).

9. **SO2**: Concentration of sulfur dioxide (µg/m^3).

10. **TEMP**: Temperature in degrees Celsius (°C).

11. **PRES**: Atmospheric pressure in hPa (hectopascals).

12. **DEWP**: Dew point temperature in degrees Celsius (°C).

13. **RAIN**: Precipitation in mm (millimeters).

14. **WSPM**: Wind speed in m/s (meters per second).

Interesting Features

1. **PM2.5 Concentration (PM2.5)**:

Significance: PM2.5 refers to fine particulate matter with a diameter of 2.5 micrometers or smaller, which can deeply penetrate the respiratory system and have adverse health effects. It is a critical quality indicator linked to respiratory diseases and air quality.

Importance: Monitoring PM2.5 levels helps assess the severity of air pollution and its potential health impacts. Analyzing trends in PM2.5 concentrations can inform policies to reduce pollution and protect public health.

2. **NO2 Concentration (NO2)**:

Significance: Nitrogen dioxide (NO2) is a common air pollutant from combustion processes, especially vehicles and industrial activities. It can irritate the respiratory system and contribute to the formation of other pollutants.

Importance: High NO2 levels indicate poor air quality and can be used to track the impact of transportation and industrial emissions. Understanding NO2 trends is crucial for urban planning and emission control strategies.

3. **O3 Concentration (O3)**:

Significance: Ozone (O3) at ground level is a secondary pollutant formed when precursor pollutants react in sunlight. While ozone is essential in the upper atmosphere, ground-level ozone can harm human health and vegetation.

Importance: Monitoring O3 levels helps understand smog formation and its effects on air quality and public health. It can also provide insights into the photochemical processes driving air pollution.

4. **Temperature (TEMP)**:

Significance: Temperature is a meteorological factor that influences various aspects of air quality, including the dispersion and chemical reactions of pollutants. Higher temperatures can accelerate the formation of ground-level ozone.

Importance: Studying the relationship between temperature and air quality can help identify seasonal variations and climate-related impacts on pollution levels. It is essential for climate change adaptation strategies.

5. **Wind Speed (WSPM)**:

Significance: Wind speed affects the dispersion and transport of air pollutants. Higher wind speeds can help disperse pollutants, while calm conditions may lead to the accumulation of pollutants in specific areas.

Importance: Understanding the role of wind speed in air quality can assist in predicting pollution episodes and designing urban layouts that promote better air dispersion. Wind patterns can also affect the distribution of pollutants across different monitoring sites.

Data Collection Period:

The dataset spans several years, from March 1, 2013, to February 28, 2017. This long-term data collection allows for analyzing air quality trends and seasonal variations.

Data Quality: real-world environmental data often comes with challenges related to missing values, outliers, and data quality. Researchers working with this dataset should be prepared to handle and preprocess the data appropriately to ensure the accuracy of their analyses.

Potential Use Cases

The "Beijing Multi-Site Air-Quality Data" can be used for various research and analysis purposes, including:

- **Air Quality Analysis**: Researchers can analyze the air quality at different monitoring sites in Beijing, identify pollution trends, and assess the impact of various factors on air quality.

- **Health Studies**: The dataset can be used to investigate the correlation between air quality and health outcomes, helping to understand the public health implications of pollution.

- **Environmental Policy**: Policymakers can use this data to formulate and evaluate policies to improve air quality in Beijing.

- **Predictive Modeling**: Machine learning models can be trained to predict air quality levels based on meteorological conditions, aiding in early warning systems.

Conclusion:

The "Beijing Multi-Site Air-Quality Data" is a valuable resource for researchers, environmentalists, and policymakers interested in understanding and addressing air quality issues in Beijing, China. It provides a rich dataset with multiple attributes and years of data, making it suitable for various analyses and applications in environmental science and public health.

Existing Analysis

1. Analysis by Alejandro Moya

URL: https://github.com/Afkerian/Beijing-Multi-Site-Air-Quality-Data-Data-Set

Tasks performed by him

Task 1: Data import, cleanup, and preprocessing

Task 2: Exploratory data analysis

Task 3: Predictive modeling

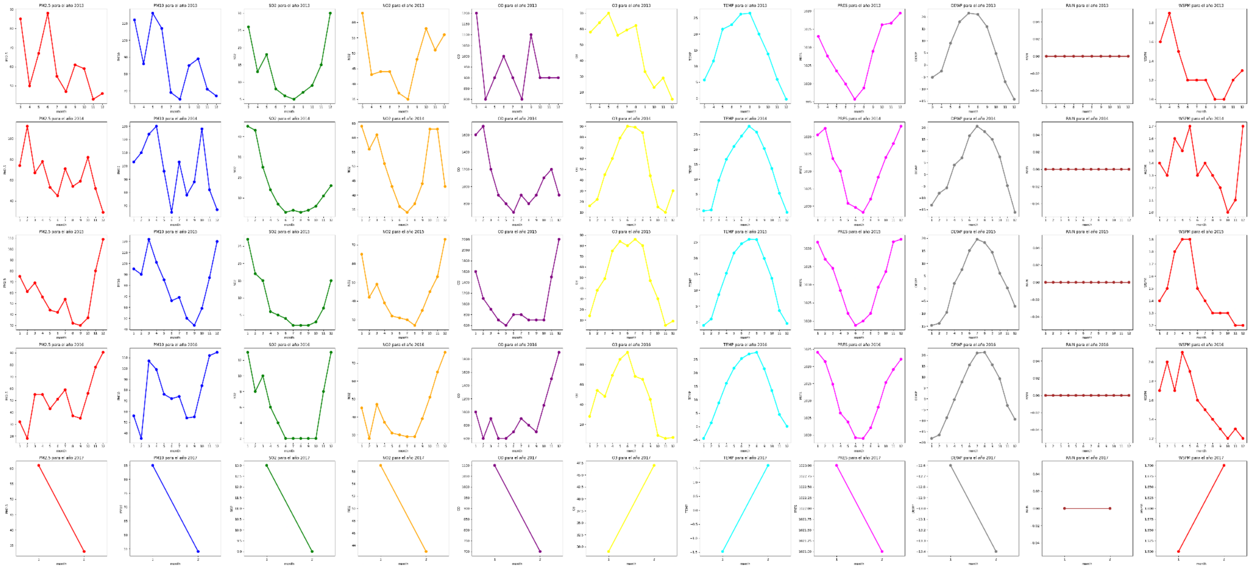


figure-2: Analysis by Alejandro Moya

2. Analysis by Rahul Sehrawat

URL: https://medium.com/@rsehrawat75/beijing-air-quality-visualisation-92c6e04ec8fd

Discoveries made in this analysis -

a) There was around 9 percent increase in air pollution from 2013-2014. Pollution decreased in 2016 if we compare it with 2015. There is around a 2 percent decrease in air pollution.
b) December is the most polluted month of the year.
c) The pollution levels remain low in the Summer months (from July to August). The slope is mostly flat
d) The pollution levels rise dramatically from March to February.
e) The second significant increase is in the month of October.
f) The station Dingling is the cleanest. It has fewer PM10 and other pollutants than Gucheng, which has the highest Pm10 particle concentration.
g) The pollution levels remain high in Winters

3. Research by Dong Li, Jiping Liu, and Yangyang Zhao

URL: https://www.mdpi.com/2073-4433/13/10/1719

Tasks done in this research-

a) Preprocessing - Wind direction, as a non-numerical type of data, needs to be converted to a numerical type of data by categorical coding. The average of the data before and following the time of the missing value is used to fill in missing values for meteorological and pollution data. Then, to eliminate the effect of numerical differences on prediction accuracy, meteorological and pollutant data were converted to the range [0, 1] by the Min-Max function as below.

b) Performance Evaluation Indices - This paper uses RMSE, MAE, R2, and IA to analyze how well the prediction models performed. The RMSE reflects the model's sensitivity to error, and the MAE reflects the stability of the model; the closer the value of both to 0, the better the prediction result. R2 represents the ability to forecast the actual data, and IA represents the distribution similarity between actual and predicted values. Both variables' values span from [0, 1]. The closer to 1, the more consistent the predicted results are with the true data distribution.

Problem Statement:

Challenges faced while preparing the dataset for analysis:

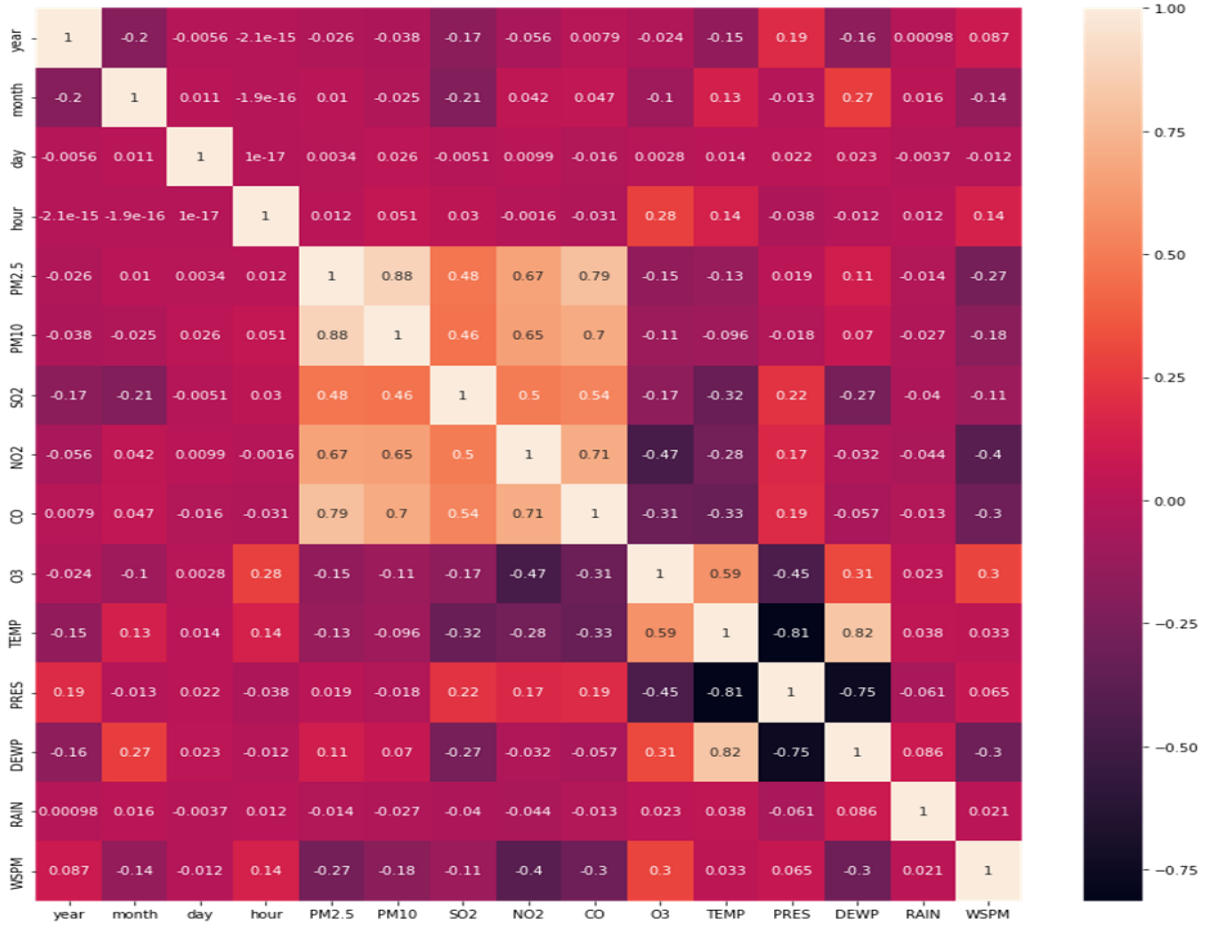
1) Properly Imputing missing values is important to retain the semantics a feature provides.

The following columns have missing values:

PM2.5, PM10, SO2, NO2, CO, and O3.

2) The rain column is highly skewed, so analysis has to be performed to see how this column can be used. Highly skewed may also imply that there may be some outliers as well, which has to be confirmed by further investigation.

Checking correlation between columns to formulate: The correlation matrix below shows a correlation between various columns of the dataset.



	year	month	day	hour	PM2.5	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	WSPM
year	1	-0.2	-0.0056	-2.1e-15	-0.026	-0.038	-0.17	-0.056	0.0079	-0.024	-0.15	0.19	-0.16	0.00098	0.087
month	-0.2	1	0.011	-1.9e-16	0.01	-0.025	-0.21	0.042	0.047	-0.1	0.13	-0.013	0.27	0.016	-0.14
day	-0.0056	0.011	1	1e-17	0.0034	0.026	-0.0051	0.0099	-0.016	0.0028	0.014	0.022	0.023	-0.0037	-0.012
hour	-2.1e-15	-1.9e-16	1e-17	1	0.012	0.051	0.03	-0.0016	-0.031	0.28	0.14	-0.038	-0.012	0.012	0.14
PM2.5	-0.026	0.01	0.0034	0.012	1	0.88	0.48	0.67	0.79	-0.15	-0.13	0.019	0.11	-0.014	-0.27
PM10	-0.038	-0.025	0.026	0.051	0.88	1	0.46	0.65	0.7	-0.11	-0.096	-0.018	0.07	-0.027	-0.18
SO2	-0.17	-0.21	-0.0051	0.03	0.48	0.46	1	0.5	0.54	-0.17	-0.32	0.22	-0.27	-0.04	-0.11
NO2	-0.056	0.042	0.0099	-0.0016	0.67	0.65	0.5	1	0.71	-0.47	-0.28	0.17	-0.032	-0.044	-0.4
CO	0.0079	0.047	-0.016	-0.031	0.79	0.7	0.54	0.71	1	-0.31	-0.33	0.19	-0.057	-0.013	-0.3
O3	-0.024	-0.1	0.0028	0.28	-0.15	-0.11	-0.17	-0.47	-0.31	1	0.59	-0.45	0.31	0.023	0.3
TEMP	-0.15	0.13	0.014	0.14	-0.13	-0.096	-0.32	-0.28	-0.33	0.59	1	-0.81	0.82	0.038	0.033
PRES	0.19	-0.013	0.022	-0.038	0.019	-0.018	0.22	0.17	0.19	-0.45	-0.81	1	-0.75	-0.061	0.065
DEWP	-0.16	0.27	0.023	-0.012	0.11	0.07	-0.27	-0.032	-0.057	0.31	0.82	-0.75	1	0.086	-0.3
RAIN	0.00098	0.016	-0.0037	0.012	-0.014	-0.027	-0.04	-0.044	-0.013	0.023	0.038	-0.061	0.086	1	0.021
WSPM	0.087	-0.14	-0.012	0.14	-0.27	-0.18	-0.11	-0.4	-0.3	0.3	0.033	0.065	-0.3	0.021	1

Figure 3: Correlation between the dataset features

The above correlation matrix has to be analyzed to draw further insights. For example, O3 and temperature have a significant correlation.

Hypotheses:

In extension to Analysis Two, which studied how pollution varies with respect to time, we would also want to study how temperature, pressure, and dew point temperature affect pollution levels.

We also want to study and analyze the effect of rain on atmospheric pollution (PM2.5, CO, etc.). Along with that, we would also like to train an ML/DL model to predict, given the state of time, station, and atmospheric condition, whether there will be high pollution or low, and also up to what will be the pollution level.

We also want to study and analyze wind direction/wind speed to see if pollution in an area is influencing other areas.

How our analysis/hypotheses would be different from existing work:

Our attempt to study how pollution in an area affects other areas is unique. Also, the existing works do not predict the pollution levels based on the atmospheric conditions, which we will do. Also, we will study the trend of wind direction and wind speed and how it influences the pollution state in other areas.